STAT679
Computing for Data Science and Statistics

Lecture 11: pandas
Pandas

Open-source library of data analysis tools
  Low-level ops implemented in Cython (C+Python=Cython, often faster)
  Database-like structures, largely similar to those available in R
  Well integrated with numpy/scipy
  Optimized for most common operations
    E.g., vectorized operations, operations on rows of a table

From the documentation: pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.
Installing pandas

Using conda:
conda install pandas

Using pip:
pip install pandas

From binary (not recommended):
http://pypi.python.org/pypi/pandas

Warning: a few recent updates to pandas have been API-breaking changes, meaning they changed one or more functions (e.g., changed the number of arguments, their default values, or other behaviors). This shouldn’t be a problem for us, but you may as well check that you have the most recent version installed.
Basic Data Structures

Series: represents a one-dimensional labeled array
  Labeled just means that there is an index into the array
  Support vectorized operations

Dataframe: table of rows, with labeled columns
  Like a spreadsheet or an R data frame
  Support numpy ufuncs (provided data are numeric)
pandas Series

By default, indices are integers, starting from 0, just like you’re used to.

But we can specify a different set of indices if we so choose.

Can create a pandas Series from any array-like structure (e.g., Python list, numpy array, dict).

Warning: providing too few or too many indices is a ValueError.
**pandas Series**

```python
1 d = {'dog': 3.1415, 'cat': 42, 'bird': 0, 'goat': 1.618}
2 s = pd.Series(d)
3 s
bird     0.0000
cat   42.0000
dog   3.1415
goat  1.6180
dtype: float64
```

Can create a series from a dictionary. Keys become indices.

```python
1 inds = ['dog', 'cat', 'bird', 'goat', 'cthulu']
2 s = pd.Series(d, index=inds)
3 s
dog   3.1415
cat   42.0000
bird  0.0000
goat  1.6180
cthulu NaN
dtype: float64
```

Index ‘cthulu’ doesn’t appear in the dictionary, so pandas assigns it NaN, the standard “missing data” symbol.
**pandas Series**

Indexing works like you’re used to and supports slices, but **not** negative indexing.

```python
s = pd.Series([2, 3, 5, 7, 11])
s[0]
```

This object has type `np.int64`

```python
s[1:3]
```

This object is another pandas Series.

```python
s[-1]
```

```
KeyError
```

```
Traceback (most recent call last)
<ipython-input-22-0e2107f91cbd> in <module>()
----> 1 s[-1]
```
**Caution:** indices need not be unique in `pandas` Series. This will only cause an error if/when you try to perform an operation that requires unique indices.
Series objects are like np.ndarray objects, so they support all the same kinds of slice operations, but note that the indices come along with the slices.

Series objects even support most numpy functions that act on arrays.
Series objects are `dict`-like, in that we can access and update entries via their keys.

**Not shown:** Series also support the `in` operator: `x in s` checks if `x` appears as an index of Series `s`. Series also supports the dictionary `get` method.

Like a dictionary, accessing a non-existant key is a `KeyError`.

**Note:** I cropped out a bunch of the error message, but you get the idea.
Entries of a Series can be of (almost) any type, and they may be mixed (e.g., some floats, some ints, some strings, etc), but they **cannot** be sequences.

Series support universal functions, so long as all their entries support operations.

Series operations require that keys be shared. Missing values become NaN by default.

To reiterate, Series objects support most numpy ufuncs. For example, `np.sqrt(s)` is valid, so long as all entries are positive.
Series have an optional name attribute. After it is set, name attribute can be changed with rename method. Note: this returns a new Series. It does not change s.name. This will become especially useful when we start talking about DataFrames, because these name attributes will be column names.
Mapping and linking Series values

Series `map` method works analogously to Python’s `map` function. Takes a function and applies it to every entry.

```
1 s = pd.Series(['dog', 'goat', 'skunk'])
2 s
0    dog
1    goat
2   skunk
dtype: object
```

```
s.map(lambda s: len(s))
```

```
0    3
1    4
2    5
dtype: int64
```
Mapping and linking Series values

```
s = pd.Series(['fruit', 'animal', 'animal', 'fruit', 'fruit'],
               index=['apple', 'cat', 'goat', 'banana', 'kiwi'])
s
apple    fruit
cat      animal
goat     animal
banana   fruit
kiwi     fruit
dtype: object
```

```
t = pd.Series({'fruit': 0, 'animal': 1})
s.map(t)
```

Series `map` also allows us to change values based on another Series. Here, we’re changing the fruit/animal category labels to binary labels.
pandas DataFrames

Fundamental unit of pandas
   Analogous to R data frame

2-dimensional structure (i.e., rows and columns)
   Columns, of potentially different types
   Think: spreadsheet (or, better, database, but we haven’t learned those, yet)

Can be created from many different objects
   Dict of {ndarrays, Python lists, dicts, Series}
   2-dimensional ndarray
   Series
Creating a DataFrame from a dictionary, the keys become the column names. Values become the columns of the dictionary.

Indices that are unspecified for a given column receive NaN.

Note: in the code above, we specified the two columns differently. One was specified as a Series object, and the other as a dictionary. This is just to make the point that there is flexibility in how you construct your DataFrame. More options: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.html
**pandas DataFrames: creating DataFrames**

Dictionary has 4 keys, so 4 columns.

```python
                           index=['Ford', 'Hoover', 'Wilson', 'Obama']),
    'PhD': {'Wilson':'Johns Hopkins'},
    'JD': {'Ford': 'Yale', 'Obama': 'Harvard'},
    'Terms': pd.Series([1,1,2,2], index=['Ford', 'Hoover', 'Wilson', 'Obama'])}
presidents = pd.DataFrame(d)
presidents
```

**Note:** Dictionary includes both text and numeric columns

By default, rows and columns are ordered alphabetically.
pandas DataFrames: row/column names

Row and column names accessible as the `index` and `column` attributes, respectively, of the DataFrame.

Both are returned as pandas Index objects.
### pandas DataFrames: accessing/adding columns

DataFrame acts like a dictionary whose keys are column names, values are Series.

```python
1 presidents['PhD']
Ford NaN
Hoover NaN
Obama NaN
Wilson Johns Hopkins
Name: PhD, dtype: object
```

Like a dictionary, we can create new key-value pairs.

```python
1 presidents['Years'] = 4*presidents['Terms']
2 presidents
```

Note: technically, this isn’t quite correct, because Ford did not serve a full term. [https://en.wikipedia.org/wiki/Gerald_Ford](https://en.wikipedia.org/wiki/Gerald_Ford)
Since the row labels are ordered, we can specify a new column directly from a Python list, numpy array, etc. without having to specify indices.

Note: by default, new column are inserted at the end. See the insert method to change this behavior: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.insert.html
### pandas DataFrames: accessing/adding columns

Scalars are broadcast across the rows.

<table>
<thead>
<tr>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
<th>Years</th>
<th>Fields Medals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Wilson</td>
<td>NaN</td>
<td>Johns Hopkins</td>
<td>2</td>
<td>Princeton</td>
<td>1</td>
<td>8</td>
</tr>
</tbody>
</table>

```python
definitions
presidents['Fields Medals'] = 0
presidents
```
Deleting columns

Delete columns identically to deleting keys from a dictionary. One can use the `del` keyword, or `pop` a key.
Indexing and selection

df.loc selects rows by their labels.
df.iloc selects rows by their integer labels (starting from 0).
Indexing and selection

<table>
<thead>
<tr>
<th></th>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
<td>0</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
</tr>
<tr>
<td>Wilson</td>
<td>NaN</td>
<td>Johns Hopkins</td>
<td>2</td>
<td>Princeton</td>
<td>1</td>
</tr>
</tbody>
</table>

```python
1 presidents['JD']
```

<table>
<thead>
<tr>
<th></th>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
</tr>
<tr>
<td>Wilson</td>
<td>NaN</td>
<td>Johns Hopkins</td>
<td>2</td>
<td>Princeton</td>
<td>1</td>
</tr>
</tbody>
</table>

```python
1 presidents[1:3]
```

<table>
<thead>
<tr>
<th></th>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
</tr>
</tbody>
</table>

```python
1 presidents.iloc[1]
```

<table>
<thead>
<tr>
<th></th>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
</tr>
</tbody>
</table>

```python
1 presidents[presidents['Terms']<2]
```

<table>
<thead>
<tr>
<th></th>
<th>JD</th>
<th>PhD</th>
<th>Terms</th>
<th>Undergrad</th>
<th>Nobels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
<td>0</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
</tbody>
</table>
Indexing and selection

Select rows by their numerical indices (again 0-indexed). This supports slices.

Note: one can also select slices with lists of column names, e.g., presidents[['JD','PhD']].
Indexing and selection

Select rows by Boolean expression.

1 presidents['JD']

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
<td>0</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
<tr>
<td>Obama</td>
<td>Harvard</td>
<td>NaN</td>
<td>2</td>
<td>Columbia</td>
<td>1</td>
</tr>
<tr>
<td>Wilson</td>
<td>NaN</td>
<td>Johns Hopkins</td>
<td>2</td>
<td>Princeton</td>
<td>1</td>
</tr>
</tbody>
</table>

1 presidents.loc['Obama']

JD | Harvard
PhD | NaN
Terms | 2
Undergrad | Columbia
Nobels | 1
Name: Obama, dtype: object

1 presidents.iloc[1]

JD | NaN
PhD | NaN
Terms | 1
Undergrad | Stanford
Nobels | 0
Name: Hoover, dtype: object

1 presidents[presidents['Terms']<2]

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ford</td>
<td>Yale</td>
<td>NaN</td>
<td>1</td>
<td>UMich</td>
<td>0</td>
</tr>
<tr>
<td>Hoover</td>
<td>NaN</td>
<td>NaN</td>
<td>1</td>
<td>Stanford</td>
<td>0</td>
</tr>
</tbody>
</table>
Indexing and selection

These expressions return Series objects.

```
presidents['JD']
```

```
presidents.loc['Obama']
```

```
presidents.iloc[1]
```

```
presidents[presidents['Terms']<2]
```
### Indexing and selection

These expressions return Series objects.

```python
presidents['JD']
```

These expressions return DataFrames.

```python
presidents[1:3]
```

More on indexing:
Arithmetic with DataFrames

```python
1 df1 = pd.DataFrame(np.random.randn(8, 4), columns=['A', 'B', 'C', 'D'])
2 df2 = pd.DataFrame(np.random.randn(5, 3), columns=['A', 'B', 'C'])
3 df1+df2
```

Pandas tries to align the DataFrames as best it can, filling in non-alignable entries with NaN.

In this example, rows 0 through 4 and columns A through C exist in both DataFrames, so these entries can be successfully added. All other entries get NaN, because $x + \text{NaN} = \text{NaN}$. 
Arithmetic with DataFrames

```
1 df = pd.DataFrame(np.random.randn(4, 2), columns=['A', 'B'])
2 df
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.331635</td>
<td>-0.500870</td>
</tr>
<tr>
<td>1</td>
<td>1.111157</td>
<td>0.293138</td>
</tr>
<tr>
<td>2</td>
<td>-0.669850</td>
<td>0.456863</td>
</tr>
<tr>
<td>3</td>
<td>0.216643</td>
<td>-0.636942</td>
</tr>
</tbody>
</table>

By default, Series are aligned to DataFrames via row-wise broadcasting. `df.iloc[0]` is a Series representing the 0-th row of `df`. When we try to subtract it from `df`, pandas forces dimensions to agree by broadcasting the operation across all rows of `df`. 

```
1 df - df.iloc[0]
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>1</td>
<td>2.442791</td>
<td>0.794009</td>
</tr>
<tr>
<td>2</td>
<td>0.661785</td>
<td>0.957734</td>
</tr>
<tr>
<td>3</td>
<td>1.548277</td>
<td>-0.136072</td>
</tr>
</tbody>
</table>
Arithmetic with DataFrames

Scalar addition and multiplication works in the obvious way. DataFrames also support scalar division, exponentiation… Basically every numpy ufunc.

DataFrames also support entrywise Boolean operations.
Arithmetic with DataFrames

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1.331635</td>
<td>-0.500870</td>
</tr>
<tr>
<td>1</td>
<td>1.111157</td>
<td>0.293138</td>
</tr>
<tr>
<td>2</td>
<td>-0.669850</td>
<td>0.456863</td>
</tr>
<tr>
<td>3</td>
<td>0.216643</td>
<td>-0.636942</td>
</tr>
</tbody>
</table>

In pandas, DataFrames support numpy-like any and all methods.

Just like numpy, direct Boolean operations are not supported.
Arithmetic with DataFrames

(values attribute stores the entries of the table in a numpy array. This is occasionally useful when you want to stop dragging the extra information around and just work with the numbers in the table.)
Arithmetic with DataFrames

DataFrames support entrywise multiplication. The $T$ attribute is the transpose of the DataFrame.

DataFrames also support matrix multiplication via the numpy-like `dot` method. The DataFrame dimensions must be conformal, of course.

**Note:** Series also support a `dot` method, so you can compute inner products.
Removing NaNs

DataFrame `dropna` method removes rows or columns that contain NaNs.

- `how='any'` will remove all rows/columns that contain even one NaN.
- `how='all'` removes rows/columns that have all entries NaN.

**axis** argument controls whether we act on rows, columns, etc.
pandas supports read/write for a wide range of different file formats. This flexibility is a major advantage of pandas.

<table>
<thead>
<tr>
<th>Format Type</th>
<th>Data Description</th>
<th>Reader</th>
<th>Writer</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>CSV</td>
<td>read_csv</td>
<td>to_csv</td>
</tr>
<tr>
<td>text</td>
<td>JSON</td>
<td>read_json</td>
<td>to_json</td>
</tr>
<tr>
<td>text</td>
<td>HTML</td>
<td>read_html</td>
<td>to_html</td>
</tr>
<tr>
<td>text</td>
<td>Local clipboard</td>
<td>read_clipboard</td>
<td>to_clipboard</td>
</tr>
<tr>
<td>binary</td>
<td>MS Excel</td>
<td>read_excel</td>
<td>to_excel</td>
</tr>
<tr>
<td>binary</td>
<td>HDF5 Format</td>
<td>read_hdf</td>
<td>to_hdf</td>
</tr>
<tr>
<td>binary</td>
<td>Feather Format</td>
<td>read_feather</td>
<td>to_feather</td>
</tr>
<tr>
<td>binary</td>
<td>Parquet Format</td>
<td>read_parquet</td>
<td>to_parquet</td>
</tr>
<tr>
<td>binary</td>
<td>Msgpack</td>
<td>read_msgpack</td>
<td>to_msgpack</td>
</tr>
<tr>
<td>binary</td>
<td>Stata</td>
<td>read_stata</td>
<td>to_stata</td>
</tr>
<tr>
<td>binary</td>
<td>SAS</td>
<td>read_sas</td>
<td></td>
</tr>
<tr>
<td>binary</td>
<td>Python Pickle Format</td>
<td>read_pickle</td>
<td>to_pickle</td>
</tr>
<tr>
<td>SQL</td>
<td>SQL</td>
<td>read_sql</td>
<td>to_sql</td>
</tr>
<tr>
<td>SQL</td>
<td>Google Big Query</td>
<td>read_gbq</td>
<td>to_gbq</td>
</tr>
</tbody>
</table>

Reading/writing files

<table>
<thead>
<tr>
<th>Format Type</th>
<th>Data Description</th>
<th>Reader</th>
<th>Writer</th>
</tr>
</thead>
<tbody>
<tr>
<td>text</td>
<td>CSV</td>
<td>read_csv</td>
<td>to_csv</td>
</tr>
<tr>
<td>text</td>
<td>JSON</td>
<td>read_json</td>
<td>to_json</td>
</tr>
<tr>
<td>text</td>
<td>HTML</td>
<td>read_html</td>
<td>to_html</td>
</tr>
<tr>
<td>text</td>
<td>Local clipboard</td>
<td>read_clipboard</td>
<td>to_clipboard</td>
</tr>
<tr>
<td>binary</td>
<td>MS Excel</td>
<td>read_excel</td>
<td>to_excel</td>
</tr>
<tr>
<td>binary</td>
<td>Msgpack</td>
<td>read_hdf</td>
<td>to_hdf</td>
</tr>
<tr>
<td>binary</td>
<td>Stata</td>
<td>read_feather</td>
<td>to_feather</td>
</tr>
<tr>
<td>binary</td>
<td>SAS</td>
<td>read_parquet</td>
<td>to_parquet</td>
</tr>
<tr>
<td>binary</td>
<td>Python Pickle Format</td>
<td>read_msgpack</td>
<td>to_msgpack</td>
</tr>
<tr>
<td>SQL</td>
<td>SQL</td>
<td>read_pickle</td>
<td>to_pickle</td>
</tr>
<tr>
<td>SQL</td>
<td>Google Big Query</td>
<td>read_sql</td>
<td>to_sql</td>
</tr>
<tr>
<td></td>
<td></td>
<td>read_gbq</td>
<td>to_gbq</td>
</tr>
</tbody>
</table>

*pandas* supports read/write for a wide range of different file formats. This flexibility is a major advantage of *pandas*.

*pandas* file I/O is largely similar to R *read.table* and similar functions, so I’ll leave it to you to read the *pandas* documentation as needed.

Table credit: https://pandas.pydata.org/pandas-docs/stable/io.html
Summarizing DataFrames

`pd.read_csv()` reads a comma-separated file into a DataFrame.

`info()` method prints summary data about the DataFrame. Number of rows, column names and their types, etc.

Note: there is a separate `to_string()` method that generates a string representing the DataFrame in tabular form, but this usually doesn’t display well if you have many columns.
Summarizing DataFrames

**head()** method displays just the first few rows of the DataFrame (5 by default; change this by supplying an argument). **tail()** displays the last few rows.

```python
baseball.head()
```

| id   | year | stint | team   | lg | g   | ab  | r   | h   | X2b | ... | rbi | sb  | cs  | bb  | so  | ibb | hbp | sh  | sf  | gidp |
|------|------|-------|--------|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| 4    | 1871 | 1     | RC1    | NaN| 25  | 120 | 29  | 39  | 11  | ... | 16.0| 6.0 | 2.0 | 2   | 1.0 | NaN | NaN | NaN | NaN  | NaN  |
| 44   | 1871 | 1     | WS3    | NaN| 32  | 162 | 45  | 45  | 9   | ... | 29.0| 8.0 | 0.0 | 4   | 0.0 | NaN | NaN | NaN | NaN  | NaN  |
| 68   | 1871 | 1     | FW1    | NaN| 19  | 89  | 15  | 24  | 3   | ... | 10.0| 2.0 | 1.0 | 2   | 0.0 | NaN | NaN | NaN | NaN  | NaN  |
| 99   | 1871 | 1     | NY2    | NaN| 33  | 161 | 35  | 58  | 5   | ... | 34.0| 4.0 | 2.0 | 3   | 0.0 | NaN | NaN | NaN | NaN  | NaN  |
| 102  | 1871 | 1     | CL1    | NaN| 29  | 128 | 35  | 45  | 3   | ... | 23.0| 3.0 | 1.0 | 1   | 0.0 | NaN | NaN | NaN | NaN  | NaN  |

5 rows × 22 columns

**Note:** R and pandas both supply `head/tail` functions, named after UNIX/Linux commands that displays the first/last lines of a file.
Comparing DataFrames

These two DataFrames *ought* to be equivalent...

...but they aren’t.
Comparing DataFrames

These two DataFrames *ought* to be equivalent...

...but they aren’t.

The problem comes from the fact that NaNs are not equal to one another.

**Solution:** DataFrames have a separate `equals()` method for checking the kind of equality that we meant above.
Comparing DataFrames

**Solution:** DataFrames have a separate `equals()` method for checking the kind of equality that we meant above.

There is a solid design principle behind this. If there are NaNs in our data, we want to err on the side of being overly careful about what operations we perform on them. We see similar ideas in `numpy` and in R.
Statistical Operations on DataFrames

Getting means of DataFrame rows/columns using numpy is possible, but tedious. Using DataFrame.mean method is a cleaner way to do the same thing. Argument picks out which axis to take means on: rows (1) or columns (0).
Of course, DataFrames also support a bunch of related functions, that work similarly: `sum`, `min`, `max`, `std`, `var` etc. All of these functions take an optional Boolean argument `skipna`. If True, NaNs are **not included** in the computation. If False, NaNs are included (which can mean either that the computation doesn’t work at all, or changes the value only slightly). More information: https://pandas.pydata.org/pandas-docs/stable/basics.html#descriptive-statistics

DataFrame's `mean` method is a cleaner way to do the same thing. Argument picks out which axis to take means on: rows (1) or columns (0).
Summarizing DataFrames

`DataFrame.describe()` is similar to the R `summary()` function. Non-numeric data will get statistics like counts, number of unique items, etc. If a DataFrame has mixed types (both numeric and non-numeric), the non-numeric data is excluded by default.

Details and optional arguments:
https://pandas.pydata.org/pandas-docs/stable/basics.html#summarizing-data-describe
Row- and column-wise functions: apply()

**DataFrame.apply()** takes a function and applies it to each column of the DataFrame.

Axis argument is 0 by default (column-wise). Change to 1 for row-wise application.
Row- and column-wise functions: `apply()`

Numpy ufuncs take vectors and spit out vectors, so using `df.apply()` to apply a ufunc to every row or column in effect ends up applying the ufunc to every element.
### Row- and column-wise functions: `apply()`

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.938898</td>
<td>2.047553</td>
<td>-0.525091</td>
</tr>
<tr>
<td>1</td>
<td>1.066293</td>
<td>-0.599466</td>
<td>-0.195606</td>
</tr>
<tr>
<td>2</td>
<td>-0.939341</td>
<td>0.022376</td>
<td>1.453082</td>
</tr>
<tr>
<td>3</td>
<td>1.114664</td>
<td>-0.408026</td>
<td>-0.811081</td>
</tr>
<tr>
<td>4</td>
<td>2.257680</td>
<td>0.280994</td>
<td>0.847329</td>
</tr>
</tbody>
</table>

We can pass positional and keyword arguments into the function via `df.apply`. `Args` is a tuple of the positional arguments (in order), followed by the keyword arguments.

```python
1 def quadratic(x, a, b, c=1):
2     return a*x**2 + b*x + c
3 df.apply(quadratic, args=(1,2), c=5)
```

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7.759325</td>
<td>13.287581</td>
<td>4.225538</td>
</tr>
<tr>
<td>1</td>
<td>8.269566</td>
<td>4.160428</td>
<td>4.647050</td>
</tr>
<tr>
<td>2</td>
<td>4.003679</td>
<td>5.045253</td>
<td>10.017612</td>
</tr>
<tr>
<td>3</td>
<td>8.471805</td>
<td>4.350433</td>
<td>4.035691</td>
</tr>
<tr>
<td>4</td>
<td>14.612481</td>
<td>5.640946</td>
<td>7.412624</td>
</tr>
</tbody>
</table>

**Note:** "`apply()` takes an argument `raw` which is `False` by default, which converts each row or column into a Series before applying the function. When set to `True`, the passed function will instead receive an ndarray object, which has positive performance implications if you do not need the indexing functionality." This can be useful if your function is meant to work specifically with Series.
Element-wise function application

This causes an error, because `apply` thinks that its argument should be applied to Series (i.e., columns), not to individual entries.
Element-wise function application

`applymap` works similarly to Python’s `map` function (and the Series `map` method). Applies its argument function to every entry of the DataFrame.
Here we have a function composition applied to a DataFrame. This is perfectly valid code, but pandas supports another approach.
Tablewise Function Application

The DataFrame `pipe` method is built for a pattern called **method chaining**. The `pipe` method has better support for passing additional arguments around than does the function composition to the right. This pattern using `pipe` is also more conducive to functional programming patterns.
Recap

Previously: basics of **pandas**
- Series and DataFrames
- Indexing, changing entries
- Function application

Next: more complicated operations
- Statistical computations
- Group-By operations
- Reshaping, stacking and pivoting
Recap

Previously: basics of Pandas
Series and DataFrames
Indexing, changing entries
Function application

Next: more complicated operations
Statistical computations
Group-By operations
Reshaping, stacking and pivoting

Caveat: Pandas is a large, complicated package, so I will not endeavor to mention every feature here. These slides should be enough to get you started, but there’s no substitute for reading the documentation.
Percent change over time

The `pct_change` method is supported by both Series and DataFrames. `Series.pct_change` returns a new Series representing the step-wise percent change.

```
s = pd.Series(np.random.randn(8))
s
0   -0.669520
1   -0.864352
2   -1.686718
3    0.014609
4   -2.199920
5   -0.505137
6   -0.403893
7  -0.358685
dtype: float64
ds.pct_change()
0   NaN
1   0.291003
2   0.951425
3  -1.008661
4 -151.589298
5  -0.770384
6  -0.200428
7  -0.111931
dtype: float64
```

**Note:** Pandas has extensive support for time series data, which we mostly won’t talk about in this course. Refer to the documentation for more.
Percent change over time

\texttt{pct\_change} operates on columns of a DataFrame, by default. Periods argument specifies the time-lag to use in computing percent change. So periods=2 looks at percent change compared to two time steps ago.

\texttt{pct\_change} includes control over how missing data is imputed, how large a time-lag to use, etc. See documentation for more detail: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.pct_change.html
Computing covariances

**cov** method computes covariance between a Series and another Series.

```
s1 = pd.Series(np.random.randn(1000))
s2 = pd.Series(0.1*s1+np.random.randn(1000))
s1.cov(s2)
```

0.1522727637202401

**cov** method is also supported by DataFrame, but instead computes a new DataFrame of covariances between columns.

```
df = pd.DataFrame(np.random.randn(1000, 4))
df.cov()
```

```
   0       1       2       3
0  1.208517 -0.515225 -0.430870  0.093096
1 -0.515225  0.673964  0.520126 -0.021969
2 -0.430870  0.520126  0.911544  0.329498
3  0.093096 -0.021969  0.329498  0.546332
```

**cov** supports extra arguments for further specifying behavior:

Pairwise correlations

**DataFrame corr** method computes correlations between columns (use `axis` keyword to change this behavior). `method` argument controls which correlation score to use (default is Pearson’s correlation.

```
dl = pd.DataFrame(np.random.randn(1000, 5),
columns=["a", "b", "c", "d", "e"])
dl.corr(method='spearman')
l
dl.corr(method='kendall')
```
Ranking data

The `rank` method returns a new Series whose values are the data ranks. For example:

```python
s = pd.Series(np.random.randn(5), index=list('abcde'))
s
# Output:
a  1.804688
b -1.203916
c  1.055365
d -0.048237
e  1.659330
dtype: float64

s.rank()
# Output:
a  5.0
b  1.0
c  3.0
d  2.0
e  4.0
dtype: float64
```

Ties are broken by assigning the mean rank to both values:

```python
s[0] = s[1] = 0
s.rank()
# Output:
a  2.5
b  2.5
c  4.0
d  1.0
e  5.0
dtype: float64
```
By default, `rank` ranks columns of a DataFrame individually.

Rank rows instead by supplying an `axis` argument.

**Note**: more complicated ranking of whole rows (i.e., sorting whole rows rather than sorting columns individually) is possible, but requires we define an ordering on Series.
Aggregating data

This command creates time series data, with rows indexed by year-month-day timestamps.

Supplying a list of functions to `agg` will apply each function to each column of the DataFrame, with each function getting a row in the resulting DataFrame.

`agg` is an alias for the method `aggregate`. Both work exactly the same.
Aggregating data

`agg` can, alternatively, take a dictionary whose keys are column names, and values are functions.

Note that the values here are strings, not functions! `pandas` supports dispatch on strings. It recognizes certain strings as referring to functions. `apply` supports similar behavior.
Aggregating data

- `df` contains mixed data types.

- `agg` (and similarly `apply`) will only try to apply these functions on the columns of types supported by those functions.

- **Note:** The `DataFrame.transform` method provides generally similar functionality to the `agg` method.

- `pandas` doesn't know how to compute a mean string, so it doesn't try.
Iterating over Series and DataFrames

Iterating over a Series gets an iterator over the values of the Series.

Iterating over a DataFrame gets an iterator over the column names.
Iterating over Series and DataFrames

The `iteritems()` method is supported by both Series and DataFrames. Returns an iterator over the key-value pairs. In the case of Series, these are (index, value) pairs. In the case of DataFrames, these are (colname, Series) pairs.

```python
for x in df.iteritems():
    print(x)
```

```python
A
('A', 0 -2.072339
1 -0.587874
2 -0.164436
3 -1.215576
4 -0.350299
Name: A, dtype: float64)

B
('B', 0 -1.282539
1 0.517591
2 1.450398
3 -0.671235
4 1.958805
Name: B, dtype: float64)

C
('C', 0 -1.241128
1 -0.394561
2 -0.975424
3 0.394053
4 0.467778
Name: C, dtype: float64)
```
Iterating over Series and DataFrames

```python
for x in df.iterrows():
    print(x)
```

DataFrame `iterrows()` returns an iterator over the rows of the DataFrame as (index, Series) pairs.
Iterating over Series and DataFrames

DataFrame `iterrows()` returns an iterator over the rows of the DataFrame as (index, Series) pairs.

**Note:** DataFrames are designed to make certain operations (mainly vectorized operations) fast. This implementation has the disadvantage that iteration over a DataFrame is slow. It is usually best to avoid iterating over the elements of a DataFrame or Series, and instead find a way to compute your quantity of interest using a vectorized operation or a map/reduce operation.
Group By: reorganizing data

“Group By” operations are a concept from databases
  Splitting data based on some criteria
  Applying functions to different splits
  Combining results into a single data structure

Fundamental object: pandas GroupBy objects
Group By: reorganizing data

```python
import pandas as pd
import numpy as np

df = pd.DataFrame({'A': ['plant', 'animal', 'plant', 'plant'],
                  'B': ['apple', 'goat', 'kiwi', 'grape'],
                  'C': np.random.randn(4),
                  'D': np.random.randn(4)})
df
```

DataFrame `groupby` method returns a `pandas groupby` object.
Group By: reorganizing data

Every `groupby` object has an attribute `groups`, which is a dictionary with maps group labels to the indices in the DataFrame.

In this example, we are splitting on the column ‘A’, which has two values: ‘plant’ and ‘animal’, so the groups dictionary has two keys.

```
1 df.groupby('A')
<
pandas.core.groupby.DataFrameGroupBy object at 0x11fe88bd0>
```

```
1 df.groupby('A').groups

{'animal': Int64Index([1], dtype='int64'),
 'plant': Int64Index([0, 2, 3], dtype='int64')}
```
Group By: reorganizing data

Every `groupby` object has an attribute `groups`, which is a dictionary with maps group labels to the indices in the DataFrame.

The important point is that the `groupby` object is storing information about how to partition the rows of the original DataFrame according to the argument(s) passed to the `groupby` method.

In this example, we are splitting on the column ‘A’, which has two values: ‘plant’ and ‘animal’, so the groups dictionary has two keys.
Group By: aggregation

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>plant</td>
<td>apple</td>
<td>0.529326</td>
<td>-0.796997</td>
</tr>
<tr>
<td>1</td>
<td>animal</td>
<td>goat</td>
<td>-0.901377</td>
<td>-0.670747</td>
</tr>
<tr>
<td>2</td>
<td>plant</td>
<td>kiwi</td>
<td>1.203032</td>
<td>1.162924</td>
</tr>
<tr>
<td>3</td>
<td>plant</td>
<td>grape</td>
<td>-0.740208</td>
<td>1.184488</td>
</tr>
</tbody>
</table>

Split on group ‘A’, then compute the means within each group. Note that columns for which means are not supported are removed, so column ‘B’ doesn’t show up in the result.

```python
df.groupby('A').mean()
```

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>animal</td>
<td>-0.901377</td>
<td>-0.670747</td>
</tr>
<tr>
<td>plant</td>
<td>0.330717</td>
<td>0.516805</td>
</tr>
</tbody>
</table>
Group By: aggregation

Here we're building a hierarchically-indexed Series (i.e., multi-indexed), recording (fictional) scores of students by major and handedness.

```python
arrs = [['math', 'math', 'econ', 'econ', 'stats', 'stats'],
        ['left', 'right', 'left', 'right', 'left', 'right']]
index = pd.MultiIndex.from_arrays(arrs, names=['major', 'handedness'])
s = pd.Series(np.random.randn(6), index=index)
s
major   handedness   1
math    left          -2.015677
         right         0.537438
econ    left          1.071951
         right        -0.504158
stats   left          1.204159
         right        -0.288676
dtype: float64
```

Suppose I want to collapse over handedness to get average scores by major. In essence, I want to group by major and ignore handedness.
Group By: aggregation

Suppose I want to collapse over handedness to get average scores by major. In essence, I want to group by major and ignore handedness.

Group by the 0-th level of the hierarchy (i.e., `major`), and take means.

We could have equivalently written `groupby('major')`, here.
**Group By: examining groups**

<table>
<thead>
<tr>
<th>major</th>
<th>handedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>math</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>econ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>stats</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>dtype: float64</td>
<td></td>
</tr>
</tbody>
</table>

`groupby.get_group` lets us pick out an individual group. Here, we're grabbing just the data from the 'econ' group, after grouping by 'major'.

```
l.s.groupby('major').get_group('econ')
```

<table>
<thead>
<tr>
<th>major</th>
<th>handedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>econ</td>
<td></td>
</tr>
<tr>
<td></td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>dtype: float64</td>
<td></td>
</tr>
</tbody>
</table>
Group By: aggregation

Similar aggregation to what we did a few slides ago, but now we have a DataFrame instead of a Series.
Group By: aggregation

Similar aggregation to what we did a few slides ago, but now we have a DataFrame instead of a Series.

Groupby objects also support the `aggregate` method, which is often more convenient.

```python
1 g = df.groupby('handedness')
2 g.aggregate(np.sum)
```

```python
1 df.groupby('handedness').mean()
```

<table>
<thead>
<tr>
<th>major</th>
<th>handedness</th>
</tr>
</thead>
<tbody>
<tr>
<td>math</td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>econ</td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>stats</td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
<tr>
<td>phys</td>
<td>left</td>
</tr>
<tr>
<td></td>
<td>right</td>
</tr>
</tbody>
</table>

```text
<table>
<thead>
<tr>
<th>handedness</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>left</td>
<td>7</td>
<td>-2.184352</td>
</tr>
<tr>
<td>right</td>
<td>7</td>
<td>-1.034337</td>
</tr>
</tbody>
</table>
```
Transforming data

From the documentation: “The transform method returns an object that is indexed the same (same size) as the one being grouped.”

Building a time series, indexed by year-month-day.

Suppose we want to standardize these scores within each year.

Group the data according to the output of the key function, apply the given transformation within each group, then un-group the data.

Important point: the result of `groupby.transform` has the same dimension as the original DataFrame or Series.
Filtering data

From the documentation: “The argument of filter must be a function that, applied to the group as a whole, returns True or False.”

So this will throw out all the groups with sum <= 2.

Like transform, the result is ungrouped.
Combining DataFrames

The `pandas.concat` function concatenates DataFrames into a single DataFrame. Repeated indices remain repeated in the resulting DataFrame. Missing values get NaN.

`pandas.concat` accepts numerous optional arguments for finer control over how concatenation is performed. See the documentation for more.
Merges and joins

*pandas* DataFrames support many common database operations
Most notably, join and merge operations

We’ll learn about these when we discuss SQL later in the semester
So we won’t discuss them here

**Important:** What we learn for SQL later has analogues in *pandas*

If you are already familiar with SQL, you might like to read this:
Pivoting and Stacking

Data in this format is usually called **stacked**. It is common to store data in this form in a file, but once it’s read into a table, it often makes more sense to create columns for A, B and C. That is, we want to **unstack** this DataFrame.
Pivoting and Stacking

The `pivot` method takes care of unstacking DataFrames. We supply indices for the new DataFrame, and tell it to turn the variable column in the old DataFrame into a set of column names in the unstacked one.

https://en.wikipedia.org/wiki/Pivot_table
How do we stack this? That is, how do we get a non-pivot version of this DataFrame? The answer is to use the DataFrame stack method.

Note: we could also construct the index set tuples using itertools.
The DataFrame `stack` method makes a stacked version of the calling DataFrame. In the event that the resulting column index set is trivial, the result is a Series. Note that `df.stack()` no longer has columns A or B. The column labels A and B have become an extra index.
Pivoting and Stacking

Here is a more complicated example. Notice that the column labels have a three-level hierarchical structure.

There are multiple ways to stack this data. At one extreme, we could make all three levels into columns. At the other extreme, we could choose only one to make into a column.
Pivoting and Stacking

Stack only according to level 1 (i.e., the animal column index).

Missing animal x cond x hair_length conditions default to NaN.
Pivoting and Stacking

Stacking across all three levels yields a Series, since there is no longer any column structure. This is often called **flattening** a table.

Notice that the NaN entries are not necessary here, since we have an entry in the Series only for entries of the original DataFrame.
Plotting DataFrames

def = pd.DataFrame(np.random.randn(1000, 3),
                  index=pd.date_range('1/1/2000', periods=1000),
                  columns=['cat', 'bird', 'goat'])

def = df.cumsum()
_ = df.plot()

cumsum gets partial sums, just like in numpy.

Note: this requires that you have imported matplotlib.

Note that legend is automatically populated and x-ticks are automatically date formatted.